Research in Practice: Understanding Significance Testing Program Evaluation

By Dale T. Griffee

A consensus is beginning to emerge that significance testing...is generally not well understood by practitioners.

Dale T. Griffee
Instructor, Department of English
South Plains College
1401 College Avenue
Levelland, TX 79336
dalegriffee@aol.com

ABSTRACT: Despite its widespread use in evaluation data analysis, statistical testing has come under persistent criticism resulting in calls for its rethinking, and even possible elimination (Carver, 1978, 1993). Saxon and Boylan issue a call "to strengthen developmental education research and to make it more accessible" (2003, p. 2). Among the types of research they consider appropriate is control group methodology which often makes use of statistical tests. This paper responds to that suggestion and seeks to explain statistical testing to state what it can and cannot tell us, and to make practical recommendations for its use.

Despite its widespread use in evaluation data analysis, Null Hypothesis Significance Testing (NHST) has come under persistent criticism resulting in calls for its rethinking, and even possible elimination (Carver, 1978, 1993; Hunter, 1997; Rozeboom, 1960; Schmidt, 1996; Schmidt & Hunter, 1997). As a result, a consensus is beginning to emerge that significance testing, with its strange backward logic and slippery terminology, is generally not well understood by practitioners (Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993). Critics also claim that significance tests are overused in a mindless and mechanistic way and interpreted in ways which cannot be supported (Rozeboom, 1997). Others (Cortina & Dunlap, 1997; Frick, 1996; Macdonald, 2002) maintain it is important to have a grasp of significance tests because they are so commonly reported in evaluation and research reports. They conclude that significance tests have some value and should be rethought rather than eliminated. A "real-life" example may help illustrate the tension in the significance testing debate.

Amanda teaches two basic writing classes, and this semester she tried an innovation in one class but not the other. Both of the classes are fairly similar, and Amanda used the same end-of-semester test in both classes. She thought the class in which she used the innovation would score higher on the test. After the final compositions were scored, Amanda found that her innovation class did, in fact, score higher than the other class. A friend told her to do a statistical significance test on the scores between the two classes. Amanda wasn't sure what it would show, but she conducted a *t*-test. Her results were $t = 4.349$, $df = 9$, $p < .0019$. Amanda went to her supervisor, showed her the results, and asked her what they meant. If you were Amanda's supervisor, what would you say?

To understand what Amanda wants to know, one must first become acquainted with significance tests, which are a family of statistical procedures for determining the probability or likelihood of achieving certain results. Because these statistical procedures involve a comparison between the actual scores and a hypothesized relationship, they are often referred to as Null Hypothesis Significance Testing; (Huberty & Pike, 1999). A null hypothesis states that there is no relationship between the results. NHST is commonly used in studies which report score data (*cf*., Boylan, Bliss, & Bonham, 1997).

Saxon and Boylan issue a call "to strengthen developmental education research and to make it more accessible" (2003, p. 2). Among the types of research they consider appropriate is that which uses control group methodology. This paper responds to their suggestion and tries to make NHST, an important tool in control group methodology, more accessible. The purpose of this paper is to explain NHST, to state what significance tests can and cannot tell us, and to make practical recommendations for their use. This paper is not a step-by-step procedure for conducting a significance test. For that, see general statistical textbooks such as Sirkin (1999, p. 217). Nor does this paper promote quantitative data at the expense of qualitative data. The following questions will be discussed:

1. What does the term "significance" mean?
2. What is a significance test?
3. What does a significance test tell us?
4. What does a significance test not tell us?
5. What recommendations can we make for using and interpreting NHST?

What Does the Term "Significance" Mean?

In ordinary language, significance indi-

cates meaning or importance. In evaluation and assessment, however, because of the ambiguity of language, we have to be more specific when we talk about evaluation results. Substantive or practical significance can be used to describe any research finding that is useful or that can have an effect on our practice. This is a judgment call, and thus there is no cut and dried test for substantive significance (Vogt, 1999, p. 219). Statistical significance applies when a significance test is run, and the low *p*-value tells us that the results were relatively rare. In this sense, statistical significance is a technical term, and it should not be confused with the ordinary meaning of the term significant.

What Is a Significance Test?

Statistics can be divided into two main categories: descriptive and inferential. Results of descriptive statistics are straightforward in the sense that they are simply summaries of scores from some instrument. An example would be the mean (average) scores from a test or a questionnaire. Inferential statistics, on the other hand, are more than summaries and allow us to take descriptive statistics one step further and use them to develop probabilities, expressed as a *p*-value. Significance tests are procedures "for determining the probability (usually at a prespecified level called alpha) of a particular result, assuming the null hypothesis to be true, given randomization and a sample of size *n*" (Shaver, 1993, p. 294). Today, significance tests are calculated mainly by computers using statistical software programs. The scores are entered into the statistical program, the type of significance test is selected (i.e., a *t*-test, ANOVA, or Pearson correlation; see Table 1), and the results are given.

The statistical result many evaluators look at first is the *p*-value. If the *p*-value is at or below a prespecified value (usually .05), then the results are judged to be statistically significant. Statistical significance is interpreted as meaning that the probability is low that we would get our results if the null hypothesis is true (Carver, 1978). So based on the low *p*-value, we decide to reject the null hypothesis which stated that there was no relationship between the scores we were comparing, and we infer that in the case of our sample there is a relationship. What that relationship is remains to be seen. All our significance test has supplied is reasonable evidence for believing a relationship might exist.

What Does a Significance Test Tell Us?

There are three terms of interest related to significance tests: probability (P), evidence (E), and the null hypothesis (H). Probability means the likelihood that something will happen, evidence is the scores from a reliable and locally validated test, and "the null hypothesis states that the experimental group and the control group are not different with respect to the ability being measured, and that any difference found between their means is due to sampling fluctuation" (Carver, 1978, p. 381). The correct formulation of NHST is P = E given H which can be read as: *p* is the probability that this evidence would arise if the null hypothesis were true.

Population means a total group of people that the researcher is interested in, for example, the population of all developmental education students enrolled in North American colleges and universities. We hypothesize that the mean scores of Group A and Group B on some measurement--for example, achievement test scores, composition ratings, GPA, or retention rates--come from the same population, and that the only difference is due to sampling error or chance. Sampling error means that since we didn't check everybody in our population, we can't be exactly sure if the ones we did pick are typical or not. We can mathematically calculate (or rather, the computer can do it for us) how often differences this large between Group A and Group B are likely to be found given the null hypothesis and other required conditions (which is the *p*-value).

If the difference between the two group scores are attributed to something other than chance, what might that something be? Maybe a treatment (perhaps a special program of teaching or classroom innovation) has made the difference but maybe not. A significance test provides no evidence as to the cause of the result (Shaver, 1993). In fact, there could be many reasons, called threats to validity, for not thinking our innovation is responsible for the increased scores. At this point we need to be a detective, not a statistician.

No paper on NHST would be complete without an understanding of assumptions. It is important to identify key assumptions associated with the statistical procedure you are using. In normal language, the word assumption means something one believes but does not know for sure. In NHST, the term "assumption" means a necessary condition. Hatch and Lazaraton (1991) say that all statistical tests have underlying assumptions which must be met in order for the interpretations to be valid and reliable. If one of the assumptions is not met, the probability is distorted, and if more than one is not met, the probability is meaningless. For example, one assumption for a *t*-test is that only two levels of one independent variable are being compared. That means a *t*-test can be used to compare two and only two sets of scores.

In addition, Shaver (1993) points out that all significance tests assume either random sampling (to provide a basis to generalize to a specific population) and/or random assignment (to provide a basis for claiming the groups are samples from the same population). If we do not use random sampling, we cannot claim that our results are true for the general population. If we do not use random assignment, we cannot be sure that our groups are not systematically different, that is that we are not comparing apples to oranges. If we compare groups that are systematically different before we engage in any teaching, say one group volunteers and the other does not, they might be different but maybe not for the reasons we think.

Cohen (1997) maintains that a significance test doesn't tell us very much. He says, if we do not reject the null hypothesis all we are saying is that the direction of the differences between Group A and Group B is uncertain. Direction in this sense means whether Group A scores are higher than Group B scores or vice vesa. On the other hand, if we reject the null hypothesis all we are saying is that we are pretty sure of the direction: For example, the mean of Group B scores are higher than the mean of Group A scores which we already knew.

To use a sports metaphor, a *p*-value at .05 or lower gives us a license to hunt. What we are hunting is an explanation of our score differences, but statistical significance itself is not that explanation (Meehl, 1978). Achieving statistical significance and rejecting the null hypothesis is a starting point, not an ending point.

By way of summary, here

Table 1
Some Commonly Used Significance Tests

Test	Some (but not all) Assumptions*
t-test	Two groups only, normal distribution
Analysis of Variance (ANOVA)	Two or more groups, normal distribution
Rank sums test	Rank-order data, nonnormal distribution
Chi squares	Frequency data
Pearson correlation	Continuous data, normal distribution
Spearman correlation	Rank-order data, nonnormal distribution

* *Note*: To find related assumptions check a text such as Hatch and Lazaraton (1991).

are four ways of expressing what p <.05 means:

- The probability of getting this evidence is 5% or less if there is no relationship.
- We can reject the null hypothesis of no relationship, which implies we believe on the basis of the evidence there is a relationship between our treatment and the test results.
- The differences we found in our scores are not likely the result of sampling error (chance).
- Chance is not a good explanation for the differences we found, even though it remains a possibility. How much of a possibility? About 5%.

What Does Significance Testing Not Tell Us?

As previously mentioned, NHST today is performed on computers using specialized statistical programs. Inputting the data is relatively easy. Unfortunately, interpreting the results is not as easy (Tryon, 1998). Following are some common, almost universal, misconceptions of NHST interpretation:

Misconception One

The *p*-value is the probability that the results were caused by chance (Macdonald, 2002). The correct conclusion, "The low *p*-value we found indicates our scores are not likely the result of chance" does not translate to, "The low *p*-value indicates the degree to which our scores were caused by chance." Amanda, from the earlier example, probably believes this misconception because the graduate student who told her to do a significance test said it would determine if her results were beyond chance. Carver (1978) says that the belief that "the *p*-value is the probability that the results were caused by chance" (p. 383) is the most common misinterpretation of significance testing. This misconception causes trouble because it leads us down the path of believing that NHST is causal: a statistical test proving that something caused something. That, in turn, leads us to think of NHST as mechanical proof. All we have to do is show $p < .05$ and we are finished. No need for further explanation of our results, no need for theory, and no need for argumentation. NHST does it all. This misconception is probably at the root of why NHST is overused and certainly at the root of why NHST is misused.

Misconception Two

The *p*-value is the probability that the null hypothesis is true. Cohen (1997) refers to this as the illusion of attaining probability. According to Cohen, even though the correct formulation is P = E/H (P is the probability of obtaining our evidence given the null hypothesis), we have a tendency to turn it around to P = H/E (P is the probability that null hypothesis is true given the evidence). However, according to Shaver (1993):

> A test of statistical significance does not indicate the probability that the null hypothesis is true or false. Rather, it provides the researcher with the information in regard to the likelihood of a result, given that the null hypothesis is true; it does not indicate the likelihood that the null hypothesis is true given a particular result. (p. 300)

Cohen believes most researchers make this mistake because they are more concerned with rejecting the null hypothesis than with the likelihood of obtaining the evidence. We are concerned with the null hypothesis because what we really want to know is whether or not our results generalize to the larger population (Thompson, 1998). In that sense, NHST does not tell us what we really want to know.

Achieving statistical significance and rejecting the null hypothesis is a starting point, not an ending point.

Misconception Three

Statistical significance means the results are important. Amanda definitely believes this to be true. The reasoning behind this misconception is that statistical significance indicates that the results are relatively rare (true), and being rare equals being important (not true; Thompson, 1996). In other words, a low *p*-value, say at .05 or lower, is taken to automatically mean the results are noteworthy, and their importance can be assumed without further discussion. When you read an evaluation report that presents *p*-values but does not discuss or interpret them, it is probably because the evaluator assumes misconception three.

Misconception Four

Statistical significance equals power. This is not true because a test of significance does not indicate the degree of effect related to a treatment. Under this misconception, a low *p*-value, say .05, is taken to mean that our score differences are powerful; this line of reasoning leads to the conclusion that an even lower *p*-value, say .01, means that our score differences are even more powerful. Researchers laboring under this illusion can be identified because they put a single asterisk next to the .05 results calling them significant, they put a double asterisk by the .01 results calling them very significant, and in some cases will even put three asterisks next to *p*-values of .001 or lower calling them highly significant. *P*-values are cut points for making a decision and do not indicate or measure the magnitude much less the importance of the result.

Misconception Five

The *p*-value is an absolute cut point (Nelson, Rosenthal, & Rosnow, 1986). A low *p*-value, usually .05 or .01, is taken as an absolute and encourages yes or no decision making. Rossi (1997) argues that it is a mistake to interpret the results of significance tests dichotomously for at least two reasons. First, .05 is arbitrary. If we have interesting results at the .056 level or even at the .06 level, is that sufficient reason to disregard them? Of course not. Some educational statisticians suggest for exploratory work a *p*-value of .10 or .15 would be acceptable (Huberty, 1987). Second, we know that if we had a larger *N* size, we would certainly reach statistical significance.

Misconception Six

Statistical significance is the probability of getting the same results if the study is replicated (Sohn, 1998). The smaller the *p*-value, it is believed, the greater the chance that if we compared another group, our results would be the same. This mistake is made because the evaluator believes the *p*-value refers to the null hypothesis and not the particular evidence on which the *p*-value is based. Since the null hypothesis could be rejected in the case of one sample (the one under consideration), it could probably be rejected with another similar sample.

Misconception Seven

Statistical significance directly reflects the probability that the converse (the research hypothesis) is true. This mistake says a *p*-value of .05 means a .95 chance of research hypothesis is true. "Even if the null hypotheses can be rejected, several other alternatives or rival hypotheses still must be ruled out before the validity of the research hypothesis is confirmed" (Carver, 1978, p. 386).

Misconception Eight

NHST encourages evaluators to look at hypotheses in a peculiar and unrealistic way. First, NHST sets us up to consider only two hypotheses when there may be many possible alternatives from which to choose (Rozeboom, 1960). More to the point, Rozeboom claims NHST treats hypothesis acceptance or rejection as if those were absolute decisions in the same sense that one accepts or rejects a piece

continued on page 32

continued from page 30

of pie for dessert, when in fact rejecting or accepting an hypothesis is a degree of believing or disbelieving. "The end product of a scientific investigation is a degree of confidence in some set of propositions, which then constitute a basis for decisions" (Rozeboom, 1960, p. 423). Thinking in absolute terms of acceptance or rejection encourages us to gloss over the hidden background and assumptions attendant in any investigation, and NHST is, to use a metaphor, one brick in the wall, not the whole wall.

What Recommendations Can We Make?

The recommendations for proper use of significance testing that follow are grouped in what I consider their range of difficulty, admittedly a subjective judgment. The first category, simple, means that the knowledge of how to implement the recommendation can be found within this paper. The second category, not as simple, means that one may have to engage in some background reading to understand and apply the recommendation and the third category, may be difficult, means that, in addition to reading, one may need a knowledgeable consultant to incorporate them appropriately.

Simple

- Insert the word "statistically" in front of "significant" when reporting results to show what kind of significance you are claiming (Carver, 1993; Thompson, 1996). Robinson and Levin (1997) think this makes journal editors into language police. Thompson (1997) acknowledges their perspectives but still believes that making clear what kind of significance is applied would be helpful, especially to new evaluators just starting their training.
- Don't mistake statistical significance for substantive significance. "Statistical significance does not mean plain-English significance" (Cohen, 1997, p. 29). Don't transform statistical significance in the results section of your paper into substantive significant in the discussion section.
- Don't confuse your *p*-value with importance (Daniel, 1998). The *p*-value is a cutoff point, not an indication of how strong your results are. Don't put a star by .05 results and two stars by .01 results. Don't indicate your results are "approaching significance" if *t*-test results are just a bit more than .05.
- Interpret the results first and the statistical significance second (Carver, 1993; Kirk, 1996). This puts your priorities in order. Again, let's remember Amanda. She gave a composition test to two of her classes, and the class that used her innovation got higher scores. Amanda should discuss those results first, and then mention later that they were statistically significant and what she thinks that means. If she finds it convenient or necessary to discuss both the results and the *p*-value together, at least she should put the results first to emphasize them.

Not As Simple

- Check the assumptions of the significance test you plan to use before you use it. What if you do not have normal distribution? What if you have frequency data instead of interval scale data? How do you check for equal variance? You should know the answers to all these questions and their implications before you decide to do a significance test. List the assumptions in your report, and briefly tell how you met them.

Amanda should discuss results first and then mention later that they were statistically significant.

- Design your study so the NHST you use is applicable and answers your research questions. You may find that you should not attempt a significance test at all. How many groups are you comparing, and is your NHST appropriate for that number? What if the groups you are comparing are not even approximately equal in size? What if one group volunteers and the other does not? Do you have random assignment or random sampling? Answers to such questions should guide the overall design of your study.
- Create an analysis section in your paper as part of the design discussion in which you state each type of analysis you plan to do, whether quantitative or qualitative, in the order in which you plan to do them. For each analysis, state the name of the analysis, why you plan to do it, and what you think it will tell you.
- Provide descriptive statistics including the *n*-size (number of persons or observations in the study) of each group, probably in table form. The descriptive statistics you provide will depend on the NHST you do, but consider giving the mean, standard deviation, minimum, maximum, skewness, and kurtosis. This provides your readers with valuable information about your groups. For example, skewness and kurtosis describe score distribution and inform readers the degree to which distribution is normal.

May Be Difficult

- Report and discuss effect size, also known as strength of association, which is an indication of how strong the effect is (APA, 2001; Thompson, 1999, 2002). Kirk (1996) discusses 40 measures of effect size. Reporting effect size is important because very small and trivial results may be statistically significant. This is because probability values are directly related to sample size. That means with a large enough number of persons in your evaluation study, virtually any association can be statistically significant (Shaver, 1993). This aspect of significance testing is often not appreciated (Mittag & Thompson, 2000). Reporting effect size helps the reader decide if your statistical significance has any practical or substantive significance.
- Build in replication. "If sampling error or chance is a reasonable threat to the generalizability of a certain result, then finding approximately the same sized result in a replication study is the best way to eliminate this threat" (Carver, 1993, p. 291). This can be done by replicating the study or by building replication into the original study. For example, you can halve the sample size and treat each half differently (such as conducting the first study before the second study) or compare the results of two groups of interest within the study population (i.e., males and females).

Conclusion

Significance tests seem to have been around forever. They are taught in most introductory statistical courses, commonly reported in the literature, and sometimes required by advisors in order to have work approved. Journal editors, professors, and reviewers often evaluate articles by the presence of statistically significant results, and besides, researchers really would like some way to judge study results.

If tests of significance are so easy to misinterpret, should we even bother with them? I vote a cautious yes, because with Abelson (1997) I suspect that if there were no significance tests, somebody would invent them.

continued on page 34

continued from page 32

That significance tests are misused is not the fault of the tests but rather of those using the tests (Cortina & Dunlap, 1997; Levin, 1998). The tests are tools and when the tools are used well and outcomes thoughtfully considered, good research emerges. I believe the real issue is causalty. We want to show that our teaching caused the results, and, although it does not show causalty, a test of significance can be one little step in the argument. Macdonald (2002) reminds us that we need to interpret the results of significance tests by taking into account the context of our study. In that sense, we are wrong to think significance tests are not subject to interpretation. Someday there may be a consensus from the education, evaluation, and measurement communities regarding what we should do about significance testing. In the meantime, if for one reason or another we decide to use a significance test, we can at least avoid the gross mistakes and misconceptions discussed in this article and can keep ourselves honest by using the recommendations provided.

References

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Abelson, R. F. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?*(pp. 117-141). Mahwah, NJ: Lawrence Erlbaum.

Boylan, H. R., Bliss, L. B., & Bonham, B. S. (1997). Program components and their relationship to student performance. *Journal of Developmental Education, 20*(3), 2-8.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*(3), 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 6*(4), 287-292.

Cohen, J. (1997). The earth is round ($p < .05$). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 21-35). Mahwah, NJ: Lawrence Erlbaum.

Cortina, J. M., & Dunlap, W. P. (1997). Logic and purpose of significance testing. *Psychological Methods, 2*(2), 161-172.

Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools, 5*(2), 23-32.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1,* 379-390.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York: Newbury House.

Huberty, C. J. (1987). On statistical testing. *Educational Researcher, 16,* 4-9.

Huberty, C. J., & Pike, C. J. (1999). On some history regarding statistical testing. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 1-22). Stamford, CT: JAI Press.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8*(1), 3-7.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Levin, J. R. (1998). To test or not to test Ho? *Educational and Psychological Measurement, 58*, 311-331.

Macdonald, R. R. (2002). The incompleteness of probability models and the resultant implications for theories of statistical inference. *Understanding Statistics, 1*(3), 167-189.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806-834.

Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 28*(4), 14-20.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41*, 1299-1301.

Robinson, D., & Levin, J. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26*(5), 21-26.

Rossi, J. S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 176-197). Mahwah, NJ: Lawrence Erlbaum.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57*(5), 416-428.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 176-197). Mahwah, NJ: Lawrence Erlbaum.

Saxon, D. P., & Boylan, H. R. (2003). Where do we go from here: An agenda for developmental education research, part I. *Research in Developmental Education, 17*(4), 1-4.

Schmidt. F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1*, 115-129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Lawrence Erlbaum.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education, 61*(4), 293-316.

Sirkin, R. M. (1999). *Statistics for the social sciences* (2nd ed.). Thousand Oaks, CA: Sage.

Sohn, D. (1998). Statistical significance and replicability: Why the former does not presage the latter. *Theory and Psychology, 8,* 291-311.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher, 26*(5), 29-32.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53*, 799-800.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology, 9*, 167-183.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25-32.

Tryon, W. W. (1998). The inscrutable null hypothesis. *American Psychologist, 53*, 796.

Vogt, W. P. (1999). *Dictionary of statistics & methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Zuckerman, M., Hodgins, H. S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. *Psychological Science, 4*, 49-53.

We are wrong to think significance tests are not subject to interpretation.

The author wishes to acknowledge K. Austin, E. Curry, G. Gorsuch, and J. Mahan for their comments on previous drafts.